The manuscript by Porz et al presents high resolution maps of the sedimentary organic carbon content and stock in the Baltic sea, derived from available observation data using DNN and MC dropout. These maps are used to initialize a numerical ocean and sediment transport model to simulate lateral fluxes of resuspended POC across the Baltic sea and its maritime boundaries. The authors conclude that the lateral transport of resuspended POC is a key factor shaping the spatial distribution of sedimentary OC and that cross boundary POC fluxes are of the same magnitude as long-term burial rates, with implications for Blue Carbon accounting.
The manuscript is well written and addresses a relevant topic at the intersection of marine geosciences and carbon budgeting. However, I have concerns regarding the ML methodology, the evaluation strategy for the OC prediction model and the interpretation of the model uncertainty. I also find the connection between ML mapping and the numerical model insufficiently exploited and I believe the manuscript would benefit from a major revision, particularly with respect to ML methodolgy and its evaluation. My major and general comments are provided below:

We thank the reviewer for their thorough review and constructive comments. We indicate below how we intend to address each comment in a revised version.

Major comments

1. The ML model is effectively a spatial interpolator, and its marginal improvement over Kriging is not adequately discussed

The DNN uses 4 input features: latitude, longitude, water depth, and distance from coast. Two of these (latitude, longitude) are spatial coordinates, meaning the model is primarily learning a nonlinear mapping from geographic location to OC content. This makes the DNN functionally similar to a spatial interpolation method, which is precisely what Kriging is designed for. The marginal improvement in prediction accuracy (RMSE: 1.75 vs. 1.81 wt%; R2: 0.664 vs. 0.656) is probably consistent with this interpretation. The DNN offers little additional predictive power beyond what spatial autocorrelation already provides.

The authors should discuss more explicitly why the DNN achieves only a marginal improvement which, in principle, is a more flexible model. I suspect it is because the feature set does not contain information beyond what spatial proximity already encodes. If the primary advantage of the ML approach is the finer-scale spatial pattern (as argued in lines 174–176), the authors need to demonstrate that this finer-scale structure is physically meaningful and not an artifact of local overfitting to nearby training points. Independent validation at the fine scale, or at least a comparison of the ML-derived patterns against known geological or bathymetric features, would help.

Furthermore, the statistical significance of the improvement over Kriging should be assessed. With a single train/test split, the observed difference in RMSE could easily fall

within the range of sampling variability. I recommend either k-fold spatial cross-validation or repeated random splits to provide confidence intervals on the performance metrics.

Finally, the paper would benefit from comparisons with simpler ML baselines (e.g., Random Forest, Gradient Boosted Trees, or even linear regression with polynomial features), if it is not out of the scope of the paper. In many cases it has been proved that the tree based models perform well for some datasets, especially tabular datasets(spatial converted to tabular) like yours. Without such baselines, it is unclear whether the DNN architecture itself contributes meaningfully to the prediction, or whether any nonlinear regression method with the same four features would perform comparably.

We agree that the advantages of the ML method have not been sufficiently shown. We plan to do a more thorough evaluation, e.g. by using k-fold cross-validation and by defining testing data using FPS in normalized feature space. Further, we aim to derive a measure of feature importance to explore the reasons for the large differences between the methods. We would also like to make the role of the evaluation protocol explicit, because we think it accounts for most of the small difference between the two methods. The FPS test set is placed deliberately far from the densely sampled regions, which makes it harder than a random hold-out of the same size: for the same model and the same training-set size, the RMSE is 1.35 wt% on a random 80/20 hold-out and 1.83 wt% on the FPS test set. Both methods are pushed towards the same limit there, so the gap between them shrinks. Where a difference can be resolved, the network is clearly ahead: by about 13 % under 5-fold cross-validation (1.27 to 1.38 against 1.59 wt%) and by about 24 % on random 80/20 hold-outs (1.51 against 1.98 wt%), with the same data, predictors and model settings. The revised manuscript will report all three protocols side by side, with the spread across folds and repetitions, so that the difference can be weighed against its sampling variability. The simpler baselines the reviewer suggests will be added as well, that is tree-based ensembles, gradient boosting and a linear reference. None of them beats the network on our data.

2. The evaluation strategy conflates model selection and model testing, and the Farthest Point Sampling scheme may not be appropriate for evaluating ML models

I have several concerns about how the model is evaluated:

(a) A cross validation technique could be useful in this case where we have less number of data to choose the best hyperparameters. Why was it not done? Is there a specific reason?

We agree that a k-fold cross-validation could make the results more robust and we will do a more thorough evaluation according to the suggestion.

(b) The Farthest Point Sampling (FPS) procedure selects test points to maximize geographic distance between them. However, the ML model uses a four-dimensional predictor space (latitude, longitude, depth, distance from coast), where "distance" has a different meaning than in geographic space. Two points far apart geographically may be similar in predictor space (e.g., two deep basins), and vice versa. The authors should justify why spatial distance is an appropriate proxy for predictor-space distance, or alternatively perform the split in normalized feature space.

It is a good suggestion to try using FPS in normalized feature space using all predictors. This could be easily implemented, and we will perform the evaluation.

(c) The additional FPS constraint, that the nearest neighbor of each test point remains in the training set, guarantees that every test point has a nearby training example. While this is practical, it systematically biases the evaluation toward favorable conditions. Both ML and Kriging will perform artificially well when the nearest data point is guaranteed to be in the training set. This may partly explain why both methods achieve similar R2 values and why neither shows strong degradation in sparse regions. The authors should report performance separately for test points in densely and sparsely sampled regions to assess whether this bias affects their conclusions.

This is a valid point and may be addressed by using FPS in normalized feature space using all predictors as suggested in (b).

3. The Monte-Carlo Dropout uncertainty estimate is not calibrated and conflates different sources of uncertainty

The standard deviation map from the Monte-Carlo Dropout procedure (Figure 3b) shows a spatial pattern that closely resembles the OC content map itself: high absolute uncertainty where OC content is high. This raises important questions:

(a) Is this pattern driven by heteroscedasticity (the model's absolute prediction variance scaling with the target magnitude), by epistemic uncertainty (insufficient training data in high-OC regimes), or by both? The paper does not attempt to disentangle these contributions. Presenting a map of relative uncertainty (coefficient of variation = standard deviation / mean) would be informative: if relative uncertainty is spatially uniform, the pattern is primarily heteroscedastic; if it varies, genuine knowledge gaps exist.

(b) The MC Dropout uncertainty is not calibrated against observed prediction errors. For the uncertainty estimate to be meaningful, the predicted confidence intervals should contain the observed values at approximately the expected rate (e.g., ~68% of observations should fall within one standard deviation). Without such calibration, the uncertainty map provides a sense of relative model disagreement across space but cannot be interpreted as a reliable quantification of prediction error.

Both points will be addressed in the revision. On (a): as implemented, the Monte-Carlo Dropout procedure samples the model parameters, not the observation noise, and the network is trained with a homoscedastic loss. The spread it produces is an estimate of epistemic uncertainty alone and contains no heteroscedastic noise term in the strict sense. What does shape the pattern is the mechanism of dropout itself. It perturbs the hidden activations multiplicatively, so the spread of the predictive samples scales roughly with the size of the prediction, and that alone yields higher absolute standard deviations where OC content is high, with no difference in data coverage behind it. We will add the map of relative uncertainty (coefficient of variation) the reviewer asks for, and show how the predictive standard deviation relates to the predicted mean on the one hand and to local sampling density on the other, which separates the two contributions. Beyond that, a variant of the network may be trained with a second output for the predictive variance under a Gaussian negative-log-likelihood loss, so that the aleatoric and the epistemic part can be reported separately.

On (b): we agree that the uncertainty estimate has to be checked against observed errors before it can be read as a prediction interval. We will report the empirical coverage of the intervals on the held-out data, together with a reliability diagram of nominal against observed coverage. The Monte-Carlo Dropout spread is much narrower than the observed error: the mean predictive standard deviation over the prediction grid is about 0.45 wt%, while the RMSE on held-out data is about 1.8 wt%, so the intervals are indeed too narrow, as the reviewer suspects. We will recalibrate the predictive distribution on held-out data with a single variance-scaling parameter, and as an alternative by isotonic recalibration following Kuleshov et al. (2018), carry the recalibrated uncertainty through into the stock estimates, and state plainly that the uncalibrated field shows relative model confidence rather than an absolute prediction error.

(c) The methods section states a dropout rate of 20% (line 81), but the hyperparameter search yields an optimal dropout rate of 10% (line 85). The dropout rate used for the final Monte-Carlo uncertainty estimation is ambiguous. Since the magnitude of dropout directly affects the spread of MC predictions (higher dropout = larger spread = larger standard deviations), this inconsistency has practical consequences for the reported uncertainties and the propagated stock estimates.

Correct, this was an oversight. An earlier version of the model used a dropout of 20%, but this was changed to 10% after introducing the hyperparameter search. This will be corrected.

(d) The MC Dropout uncertainty captures only one component of the total prediction uncertainty. Structural uncertainties, such as the choice of features, network architecture, or training data composition, are not included. Given that the ML and Kriging maps differ by more than 5 wt% in some areas (Figure 3d), the structural uncertainty may substantially exceed the MC Dropout uncertainty in those regions.

We agree that Monte-Carlo Dropout only captures the uncertainty from the parameters of one architecture. We plan to address this by including several model classes, namely the network ensemble, the tree-based ensemble, the combined model and Kriging, and their per-pixel spread gives a measure of structural uncertainty; the total will be reported as the combination of the Monte-Carlo and the structural term. The contribution of the network initialisation can be separated out from the spread across ensemble members. We expect this to confirm what the reviewer suspects, that structural uncertainty exceeds the Monte-Carlo Dropout uncertainty wherever the ML and Kriging maps differ most, and we will say so where the uncertainty maps are shown.

4. The connection between the ML-based OC mapping and the numerical transport modelling is underexploited

The paper presents the OC mapping and the lateral transport simulation as largely independent analyses, connected only by the initialization of the sediment model with the ML-derived OC map. However, a more integrated analysis could substantially strengthen the paper's central conclusion that lateral transport controls OC distribution:

(a) The OC map from the ML model is used to initialize the transport model, but no sensitivity analysis is presented to show how the results would differ if the Kriging map were used instead. Given the pronounced differences between the two maps (Figure 3d, >5 wt% in some areas), the simulated lateral fluxes could differ substantially. If the flux results are robust to the choice of initialization map, this would strengthen confidence in the transport conclusions. if not, it would highlight a key source of uncertainty.

It is an interesting idea to repeat the simulations with the Kriging map. Since the simulations are very costly in terms of computational demand, we may simulate part of the period (e.g. one selected year with maximum transport over the entire period) with initialization according to the Kriging solution to get a sense of this effect.

(b) The simulated near-bottom transport fields could be used as additional features in the ML model to directly test whether lateral transport explains observed OC patterns. If including simulated transport improves prediction accuracy and shows high feature importance (e.g., via SHAP values, could be included if it is not too much out of the scope), this would provide model-based evidence for the paper's main claim, rather than relying on the indirect argument that flux magnitudes are "of the same order" as burial rates.

We think that using the simulated near-bottom transport as an additional predictor of OC concentrations would not be very meaningful, since the result of the ML model was used to initialize the sediment fields of the transport model, so feeding the results of this model back into the ML prediction would result in a kind of autoregressive feedback that would only amplify existing biases. However, we do agree that additional analysis of the transport patterns could be used to strengthen the argument that lateral transport is important for OC accumulation, and we will add such analysis to the revision.

5. The sediment transport model lacks validation for the quantities of interest

The hydrodynamic model is validated against MARNET monitoring data for salinity and current velocity (Appendix B), which is appropriate for the hydrodynamic component. However, the sediment transport module, which produces the central results on lateral POC fluxes , has no validation against observed suspended sediment or POC concentrations. The authors acknowledge this (lines 416–424) but proceed to draw quantitative conclusions about POC fluxes at the MtC/yr scale. This is a significant limitation that should be discussed more prominently, not only in the limitations section but also when presenting the flux results.

The POC sediment class is parameterized identically to inorganic silt (Table A1: same settling velocity, critical shear stress, and erosion rate), which is a substantial simplification. OC-rich aggregates typically have lower density than mineral silt, affecting both settling and erodibility. No sensitivity analysis is presented for these parameters. Given that the simulated fluxes span orders of magnitude (Figure 6), even moderate changes in critical shear stress or settling velocity could substantially alter the flux magnitudes and patterns.

We agree that the transport formulation may be sensitive to selected parameters. However, we do note that the modelling system has been used in the past to simulate sediment transport in the Baltic Sea, with reasonable magnitudes when compared to the few SPM measurements available (Porz et al., 2021, 2023). The high computational demand currently prohibits a full sensitivity analysis of all sediment parameters. However, we may perform a few additional single-year simulations of the most critical parameters (such as settling velocity and erosion rate) in order to gauge the uncertainty in those parameters.

General comments

L17–18: The term "maritime boundaries" in the abstract is ambiguous. It should be clarified that these refer to boundaries between countries' maritime zones (EEZs, territorial seas, etc.), and the relevance for national carbon budgets should be stated explicitly.

We agree that this may be confusing and will attempt to clarify the phrasing in the abstract. It is difficult at this stage to make definitive statements on relevance for national carbon budgets, since it is not yet clearly defined which parts of a marine area (such as EEZ and territorial sea) shall count towards a country's national carbon inventory. But we agree that this point could be discussed more explicitly, and we will do so in the revised version.

L60–63: The three OC datasets are combined without specifying how overlapping measurements are handled. Were duplicate locations averaged, preferentially selected

from one dataset, or retained as separate points? This could affect both the spatial density weighting and the model training.

Overlapping measurements contained exactly the same data, since the datasets themselves partially contained the same data, so averaging was not needed. This will be clarified in the text.

L65: The four features are described as "input features (predictor variables)," but latitude and longitude are spatial coordinates rather than predictive features in the traditional ML sense.

We argue that latitude and longitude can be used as predictors, since zonal and meridional variations of seabed properties (including OC contents) are expected, especially in areas with large geographic climatic variation such as the Baltic. This explanation will be added.

L67–68: The spatial density weighting is described as "weighted inversely with the number of points within a defined radius," but the radius is not specified here (it appears later as the hyperparameter "spatial weighting radius: 0.25 degrees"). This should be described more completely, including how extreme weights for isolated points in sparse regions are handled.

We agree with the suggestion, this will be described in more detail.

L81 vs. L85: The dropout rate is stated as 20% in the methods description but 10% in the optimal hyperparameters. Please clarify which value is used for the final MC Dropout runs.

Correct, this was an oversight. An earlier version of the model used a dropout of 20%, but this was changed to 10% after introducing the hyperparameter search. This will be corrected.

L89–92: The heuristic constraint in FPS (nearest neighbor of each test point remains in training) should be discussed in terms of its potential bias on the evaluation metrics, as noted in Major Comment 2c.

This is a valid point and may be addressed by using FPS in normalized feature space (without the additional constraint) using all predictors as suggested.

L110–114: The stock calculation chain (OC content x dry bulk density x thickness) involves multiple intermediate maps, each with their own uncertainties. The propagation of uncertainty through this chain should be described more formally, and the assumption of Gaussian error propagation should be justified.

We will check for normal distribution and explain the error propagation used more clearly.

L133: The simulation period 2010–2015 is relatively short. Given the high interannual variability shown in Figure 7, the authors should discuss whether 6 years is sufficient to derive representative mean fluxes and whether the choice of period (which includes the exceptionally strong 2014/2015 MBI) biases the results.

There is strong evidence that the bottom circulation and inflows in the Baltic Sea have remained relatively stable throughout the last decades (Mohrholz, 2018; Jędrasik and Kowalewski, 2019). Nevertheless, our results do show interannual variability in POC flux for some regions, reflecting the interannual variability in hydrodynamics. Unfortunately, the high computational demand prohibits multi-decadal simulations at this stage, but this may be necessary for longer-term estimates (as noted in the discussion section). The model period was selected to capture the strong inflow event of 2014/2015, though its effect is not immediately obvious in the computed POC fluxes. We may add a measure of simulated flux stability to discuss this further.

L149–150: The claim that the OC sediment class represents "recalcitrant OC" is not supported by the data description. The OC measurements used for mapping (Section 2.1) appear to be total OC from surface sediments, with no indication that labile and refractory fractions were distinguished. This should be clarified.

The recalcitrant fraction has not been separated, as we assumed the contribution of the labile fraction to be relatively minor. This assumption will be qualified in the revised version.

L169–170: The statement that "the ML method achieves slightly more accurate predictions" should be tempered. A reduction in RMSE from 1.81 to 1.75 wt% (3.3% improvement) based on a single train/test split may not be statistically significant. The authors should either provide confidence intervals or refrain from drawing conclusions about relative performance.

Thank you for pointing this out. We will test the robustness of this using k-fold cross-validation.

L174–176: The claim that the ML method provides "much more detailed and finer-scale distribution pattern" is presented as an advantage, but it could equally reflect overfitting to local training data. Without independent validation at the fine scale, the physical reality of these patterns cannot be assessed.

Validation at the fine scale is difficult due to the distribution of data points in space. We hope that a more thorough validation using k-fold cross validation and normalized FPS along all dimensions will give a clearer picture. Alternatively, this conclusion will need to be rephrased.

L210: "top meter" ? should this be "top 10 cm"?

Yes, this will be corrected, thank you.

L271–275: The comparison with Parameswaran et al. (2025) is based solely on the basin-mean OC content (3.06% vs. 3.03%). Agreement in the mean does not imply agreement in spatial patterns. To support the claim that four features suffice, the authors should compare spatial patterns at the regional level, particularly in data-sparse regions where the 139-feature model might capture process information that geographic proxies cannot.

We agree with the comments. We will perform a quantitative comparison to Parameswaran et al. (2025) to test this claim.

L277–281: The authors argue that additional features may introduce biases from hydrodynamic model outputs. While this is a valid concern, it applies equally to the numerical transport model used later in the paper, which also relies on prescribed bottom roughness and other poorly constrained parameters. This argument is therefore somewhat self-contradictory.

We do not agree that this argument is contradictory, since we do not use our hydrodynamic model outputs for the ML prediction for precisely this reason (see response to 4b). For the hydrodynamic model itself, it is necessary to choose suitable parameters, and we aim to perform sensitivity tests to check the sensitivity to the most important ones (e.g., settling speed and erosion rate) based on a range of values recommended in existing literature.

L376–378: The attribution of the stock difference with Parameswaran et al. to porosity assumptions is plausible but not tested. A direct sensitivity analysis — recomputing the stock with the Martin et al. (2015) global porosity map — would strengthen this claim.

Agreed, the quantitative comparison to Parameswaran et al. (2025) should help to clarify this claim.

L389–395: The discussion of pre-Littorina vs. post-Littorina OC content is important but would benefit from acknowledging that anthropogenic perturbations over the last ~100 years (eutrophication, soil erosion, land-use change) may have also induced significant shifts in OC burial that are distinct from the Littorina transgression signal.

While more recent shifts may also have altered OC burial, we consider them relatively minor compared to those during the early Holocene (at least, we do not see a comparable recent shift from sediment cores). Nevertheless, we will add this aspect to the discussion and account for recent changes using depth-resolved OC content information (Langley et al., 2026).

L400–406: The distinction between "stable" and "fresh" POC is central to the transport modelling but is not operationally defined. How was the recalcitrant fraction identified or quantified from the available OC measurements? If the surface OC data represent total OC, the initialization of the transport model with total OC as "recalcitrant POC" requires justification.

The “fresh” fraction has not been separated. Samples are taken mainly with grab samplers and are assumed to represent the top ~10 cm, while the fresh fraction is restricted mainly to the seafloor surface. Therefore, we assume that the sediment OC field initialized from the OC map represents mainly the stable fraction, with minor contribution of the fresh surface material. This will be better explained in the revised version. We may also include a 1-year simulation separating POC into fractions of different labilities based on a biogeochemical model output to estimate the uncertainty.

Section 3.3 / Figures 6–8: It would be valuable to assess sensitivity of the simulated fluxes to the choice of OC initialization map (ML vs. Kriging). If the flux patterns are robust, this strengthens the transport conclusions; if not, it highlights a key uncertainty.

Agreed, we aim to perform a sensitivity test using the Kriging map as an initial field for the transport model.

Overall, it is a very nice paper and has a lot of potential. This could be very useful for carbon budgeting in the Baltic which has a lot of uncertainties, although there is a lot of data points compared to other marine regions in the world.

We thank the reviewer for their thorough review and constructive comments.

References

Jędrasik, J. and Kowalewski, M.: Mean annual and seasonal circulation patterns and long-term variability of currents in the Baltic Sea, J. Mar. Sys., 193, 1–26, doi:10.1016/j.jmarsys.2018.12.011, 2019.

Kuleshov, V., Fenner, N., and Ermon, S.: Accurate uncertainties for deep learning using calibrated regression, in: International conference on machine learning, PMLR, 2796–2804, 2018.

Langley, B., Burdett, H. L., Cameron, K., Juul-Pedersen, T., Rouillard, A., Slaymark, C., and Kamenos, N. A.: Hotspots of Arctic and sub-Arctic marine sediment organic carbon are dominated by the Baltic, Barents and Chukchi Seas, Comm. Earth Environ., 7, 529, doi:10.1038/s43247-026-03720-8, 2026.

Mohrholz, V.: Major Baltic Inflow Statistics – Revised, Front. Mar. Sci., 5, 384, doi:10.3389/fmars.2018.00384, 2018.

Parameswaran, N., González, E., Burwicz-Galerne, E., Braack, M., and Wallmann, K.: NN-TOC v1: global prediction of total organic carbon in marine sediments using deep neural networks, Geosci. Model Dev., 18, 2521–2544, doi:10.5194/gmd-18-2521-2025, 2025.

Porz, L., Zhang, W., and Schrum, C.: Density-driven bottom currents control development of muddy basins in the southwestern Baltic Sea, Mar. Geol., 438, 106523, doi:10.1016/j.margeo.2021.106523, 2021.

Porz, L., Zhang, W., and Schrum, C.: Natural and anthropogenic influences on the development of mud depocenters in the southwestern Baltic Sea, Oceanologia, 65, 182–193, doi:10.1016/j.oceano.2022.03.005, 2023.